XoFTR: Cross-modal Feature Matching Transformer

Önder Tuzcuoğlu[1,3] Aybora Köksal[1,3] Buğra Sofu[4] Sinan Kalkan[2,3] A. Aydın Alatan[1,3]
[1] Dept. of Electrical and Electronics Eng. [2] Dept. of Computer Eng.
[3] Center for Image Analysis, Middle East Technical University, Ankara, Turkey
[4] ROKETSAN Inc., Ankara, Turkey
[1,2,3]{tuzcuoglu.onder, aybora, skalkan, alatan}@metu.edu.tr, [4]bugra.sofu@roketsan.com.tr

Abstract

We introduce, XoFTR, a cross-modal cross-view method for local feature matching between thermal infrared (TIR) and visible images. Unlike visible images, TIR images are less susceptible to adverse lighting and weather conditions but present difficulties in matching due to significant texture and intensity differences. Current hand-crafted and learning-based methods for visible-TIR matching fall short in handling viewpoint, scale, and texture diversities. To address this, XoFTR incorporates masked image modeling pre-training and fine-tuning with pseudo-thermal image augmentation to handle the modality differences. Additionally, we introduce a refined matching pipeline that adjusts for scale discrepancies and enhances match reliability through sub-pixel level refinement. To validate our approach, we collect a comprehensive visible-thermal dataset, and show that our method outperforms existing methods on many benchmarks. Code and dataset at https://github.com/OnderT/XoFTR.

1. Introduction

Matching local features across different views of a 3D scene is a fundamental step for e.g. visual camera localization [11, 64], homography estimation [24], and structure from motion (SfM) [66]. Matching features between visible-thermal images is a special case in image matching. Unlike visible images, thermal infrared (TIR) images are robust against adverse light and weather conditions such as rain, fog, snow, and night [19, 40]. However, visible-TIR image matching faces challenges due to differences in texture characteristics and nonlinear intensity differences between the thermal and visible spectra, stemming from distinct radiation mechanisms: Thermal images depict thermal radiation, while visible images capture reflected light [47]. TIR cameras also typically have lower resolution and field of view [40, 71], affecting matching performance.

To match TIR-visible images, many hand-crafted [10, 29, 35, 37, 45] and learning-based [1, 8, 15, 17, 55] methods have been proposed. Despite the promising results reported, performances across different viewpoints, scales, and poor textures have been sub-optimal. Learning-based methods for visible image matching have addressed many of these challenges but often overlook extreme modality differences in thermal-visible pairs [14, 38, 49, 69, 89].

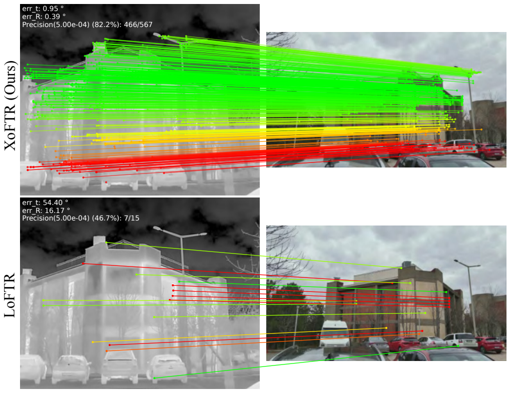

Figure 1. Our XoFTR provides significant improvements over LoFTR [69] on visible and thermal image pairs. Only the inlier matches after RANSAC are shown, and matches with epipolar error below 5×10^{-4} are drawn in green.

To address this gap, our study endeavors to extend the methods for visible image matching advancements to the visible-TIR matching problem, choosing the LoFTR network [69] as our baseline model. LoFTR, recognized for its effectiveness in matching through the use of self-attention layers and a correlation-based refinement at the subpixel level, is robust in challenging scenarios, partially also due to the training on the MegaDepth dataset [46]. However, visible-only training sets limit performance in visible-TIR matching. To address this, we propose a two-stage approach adapting (i) masked image modeling (MIM) pre-training and (ii) fine-tuning with augmented images. Inspired by PMatch's MIM [89], our strategy introduces the model to TIR-visible differences, enhancing overall performance. For fine-tuning, we use a robust pseudo-TIR image augmentation method to adapt to modality-induced variations, extending the cosine transform [27, 83].

As highlighted in prior work [38], LoFTR faces challenges with scale differences, often in thermal imaging due to lower resolutions and narrower fields of view. Inspired by AdaMatcher [38], we introduce one-to-one and one-to-many matches at 1/8 the original resolution during coarse matching and propose a fine matching pipeline that upscales these matches to 1/2 scale with a customized decoder for both pre-training and fine-tuning, enhancing visible-thermal matching. Re-matching at 1/2 scale filters low-confidence matches, improving reliability of textural structures. Matches are then refined at sub-pixel level using a regression mechanism, preventing a point in one image matching with multiple points in the other.

The absence of a suitable urban visible-thermal benchmark has led us to develop a new dataset for evaluating our method, covering a wide range of viewpoint differences and weather conditions (sunny and cloudy) for comprehensive evaluation. We also evaluate our method's homography estimation performance on publicly available datasets, demonstrating that it surpasses strong baselines, achieving state-of-the-art performance.

Contributions. Our main contributions are as follows:

- We introduce a novel two-stage training approach for visible-thermal image matching, addressing dataset scarcity by leveraging masked image modeling pre-training and fine-tuning with augmented images.
- We propose an innovative fine matching pipeline suitable for the pre-training phase of visible-thermal image matching, enabling one-to-many matching and ensuring reliable texture matching at a reduced scale of 1/2.
- We curate a novel challenging visible-TIR image matching dataset covering various viewpoint differences and weather conditions.
- Through rigorous experiments, we show that our approach outperforms strong baselines, achieving state-of-the-art results in visible-thermal image matching.

2. Related Work

Local Feature Matching. Detector-based methods for local feature matching can be categorized into handcrafted and learning-based approaches. Following the success of [9, 12, 54, 61], handcrafted methods used to be popular before the rise of deep learning based techniques [25, 70, 84]. Deep networks such as Superpoint [20] and R2D2 [59] introduced self-supervised models and joint learning techniques for improved keypoint detection and descriptor discrimination. Graph-based methods such as SuperGlue [63] and LightGlue [49] improved efficiency and matching accuracy through graph neural networks and optimized algorithms.

Detector-based approaches often struggle in low-texture areas, a problem which is mitigated in detector-free end-to-end learning-based methods [26, 60, 72, 73, 87]. The use of Transformer in detector-free matching provided state-of-the-art results [14, 38, 69, 80, 89]. A prominent example, LoFTR [69], utilizes a Transformer architecture for local image feature matching, generating matches from coarse to fine, especially in low-texture regions. Other more recent examples include AdaMatcher [38] and PMatch [89]. AdaMatcher tackles large-scale and viewpoint variations with an innovative feature interaction module and adaptive matching for precise patch-level accuracy. PMatch [89] redefines dense geometric matching through a novel approach to masked image modeling, a cross-frame transformer, and a unique loss function that improves performance in textureless areas.

Method	Year	Multimodal	Multiview
LoFTR [69]	2021	×	✓
DKM [26]	2023	×	✓
Shape-Former [15]	2023	✓	×
MIVI [23]	2023	✓	×
AdaMatcher [38]	2023	×	✓
PMatch [89]	2023	×	✓
XoFTR	**Ours**	✓	✓

Table 1. A comparative study of our approach with prior work.

While all these methods attain high performances in visible imagery, their application to multimodal visible-thermal pairs is limited. Some studies focus to fill this gap, ranging from handcrafted techniques [10, 29, 36, 50, 51] to learning-based solutions [1, 3, 7, 17, 55, 76]. ReDFeat [18] recouples detection and description constraints with a mutual weighting strategy, increasing the training stability and the performance of the features. Shape-Former [15] and MIVI [23] represent advanced matching methods for multimodal image pairs, emphasizing feature matching and structural consensus.

As summarized in Tab. 1, proposed work distinguishes itself from previous approaches by supporting both multimodality and multiview simultaneously. XoFTR is robust across various angles and scales, as well as textures on objects in images of different modalities.

Unsupervised Pre-training in Vision. Following BERT [22] and GPT [58] in NLP, unsupervised pre-training (UPT) has become widely used in computer vision, notably with DINO [13]. Inspired by context autoencoders [57], denoising autoencoders [75], and masked language modeling as a UPT task in BERT, many studies [6, 86, 88] has been introduced MIM as a UPT for learning useful representations of images by predicting masked image regions. Approaches using MIM explored different masking and UPT strategies: e.g., the Masked Autoencoder (MAE) [34] focuses on partially observed patches, whereas SimMIM [81] operates by random selection from fully observed patches.

Despite the unprecedented success of UPT strategies in tasks using RGB imagery, the research on UPT in multi-modal settings has been sparse and only very recent. Multi-MAE [4] is one of the few attempts that enhances cross-modal learning by reconstructing masked patches from different modalities, improving task performance without needing specific multi-modal datasets. Complementary Random Masking [68] targets RGB-Thermal segmentation with unique masking and self-distillation approach, reducing modality dependency. Additionally, a multi-modal transformer [16] employs masked self-attention for efficient learning with incomplete multi-modal data.

Our presented approach differs from the aforementioned studies on multi-modal UPT by introducing a scheme to adapt the paired pre-training method of PMatch [89] to the network structure of LoFTR for Visible to Thermal image applications. This method is adaptable for both pre-training and fine-tuning stages.

RGB - Thermal Image Conversion. The literature discusses hand-crafted and learning-based methods for RGB to thermal image conversion. One of the simplest yet noteworthy hand-crafted approaches is the cosine transform, introduced by Fookes et al. [27] and further explored by Yaman and Kalkan [83]. This method is simple and computationally light but produces images lacking true thermal characteristics.

To generate more realistic thermal imagery, several studies have introduced learning-based methods for RGB to TIR, e.g., using Generative Adversarial Network (GAN) based unpaired image to image translation [21, 42, 85, 90]. The central goal in such approaches is to understand the correlation between RGB an TIR images and simulate thermal image as realistic as possible. However, GANs may generate artifacts, if test images differ significantly from training data. To mitigate this, contrastive learning [56] and dual contrastive learning [30] have been proposed. A multi-domain translation network introduced by Lee [44], and its edge-preserving modification [43], use separate vectors for content and style, aiding in domain translation and preserving edges, respectively, even in the absence of annotated TIR datasets.

Despite these recent advancements, generative networks for TIR conversion still produce outputs with assumptions and artifacts, leading to pixel inaccuracies [19]. Therefore, the cosine transform method is preferred for the purposes of this work due to its reliability.

3. Methods

Our method XoFTR utilizes a ResNet-based [33] CNN for multi-scale feature extraction from visible and thermal images, integrating three modules: coarse-level and fine-level matching modules, and a sub-pixel refinement module, for precise image match predictions at multiple resolutions. Starting with feature extraction, the approach progresses through coarse and fine-level matching to determine image feature correspondence across scales, and concludes with sub-pixel refinement for accurate match localization. We introduce a new paired masked image modeling (MIM) method of PMatch [89] for semi-dense matching, pre-training with real image pairs and fine-tuning with pseudo-thermal images created from visible image datasets through cosine transform augmentation. An overview of the proposed method is presented in Fig. 2.

3.1. Coarse-Level Matching Module

The coarse-level matching module aims to predict matches at a $1/8$ scale of the original image resolution using coarse-level features $F^A_{1/8}, F^B_{1/8}$ derived from the CNN backbone. Unlike the original LoFTR architecture, which employs one-to-one assignment for coarse-level match predictions, we adopt the many-to-one/one-to-many/one-to-one assignment strategy proposed in AdaMatcher [38]. This approach addresses feature inconsistency caused by large-scale or viewpoint variations common in visible-thermal image matching, eliminating the need for manual methods such as image cropping.

LoFTR Module: We directly adopted the LoFTR module [69], which consists of linear self- and cross-attention [41] blocks, to correlate the feature maps $F^A_{1/8}$ and $F^B_{1/8}$, providing refined feature maps denoted as $\hat{F}^A_{1/8}$ and $\hat{F}^B_{1/8}$.

Matching Layer: Given $\hat{F}^A_{1/8}$ and $\hat{F}^B_{1/8}$, firstly, the similarity matrix $\mathcal{S}$ is calculated as:

$$\mathcal{S}(i,j) = \frac{1}{\tau} \cdot \left\langle \mathrm{Linear}(\hat{F}^A_{1/8}(i)), \mathrm{Linear}(\hat{F}^B_{1/8}(j)) \right\rangle, \tag{1}$$

where $\mathrm{Linear}(\cdot)$ is linear layer, i and j indices of features in feature map $\hat{F}^A_{1/8}$ and $\hat{F}^B_{1/8}$, and $\langle \cdot, \cdot \rangle$ stands for the inner product. Inspired by [38], coarse-level matching probability matrices are obtained by:

$$\begin{aligned} \mathcal{P}^0(i,j) &= \mathrm{Softmax}\left(\mathcal{S}\left(i,\cdot\right)\right)_j, \\ \mathcal{P}^1(i,j) &= \mathrm{Softmax}\left(\mathcal{S}\left(\cdot,j\right)\right)_i. \end{aligned} \tag{2}$$

From the matching probability matrices $\mathcal{P}^0$, we select pairs (i,j) as matches when the corresponding confidence values are higher than a threshold value θ_c and than any other element along its rows. Similarly, for $\mathcal{P}^1$, indexes higher than the threshold and other elements in the column were selected as match predictions. We represent coarse-level match predictions as:

$$\begin{aligned} \mathcal{M}_c = \quad & \{(\tilde{i},\tilde{j}) \mid \mathcal{P}^0(\tilde{i},\tilde{j}) = \max_k \mathcal{P}^0(\tilde{i},k), \mathcal{P}^0\left(\tilde{i},\tilde{j}\right) \geq \theta_c\} \bigcup \\ & \{(\tilde{i},\tilde{j}) \mid \mathcal{P}^1(\tilde{i},\tilde{j}) = \max_k \mathcal{P}^1(k,\tilde{j}), \mathcal{P}^1\left(\tilde{i},\tilde{j}\right) \geq \theta_c\}. \end{aligned} \tag{3}$$

3.2. Fine-Level Matching Module

Given the coarse-level match predictions ($\mathcal{M}_c$), FLMM is employed to attain matches at the $1/2$ scale of the origi-

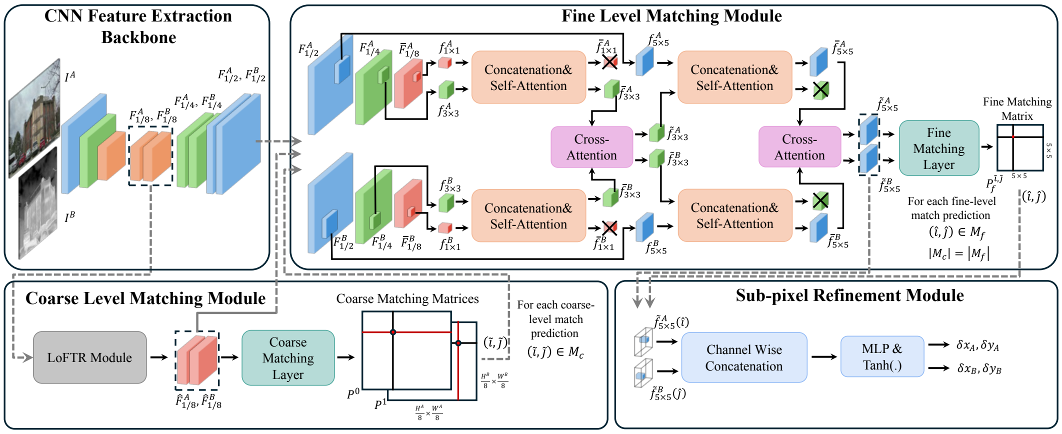

Figure 2. **Overview of the proposed method.** XoFTR consists of four modules: (1) A CNN backbone which extracts features at scales of $1/8$, $1/4$, and $1/2$. (2) The coarse-level matching module (CLMM), which correlates the features and creates coarse-level match predictions (at $1/8$ scale), allowing one-to-one and one-to-many assignment. (3) The fine-level matching module (FLMM), which re-matches coarse-level match predictions at $1/2$ scale and creates fine-level match predictions, filtering low-confidence matches. (4) The sub-pixel refinement module (SPRM) for refining fine-level match predictions at the sub-pixel level with a regression mechanism, preventing a point in one image from matching with multiple points in the other image.

nal image resolution. For this purpose, we designed a customized decoder architecture that permit it to be used both in the pre-training phase and to carry the information processed by the LoFTR module to upper layers, enhancing fine-level visible-thermal matching ability. Furthermore, matches at the 1/2 scale undergo reassessment based on confidence values, enabling the selection of texturally more reliable matches.

In the decoder structure, firstly, we concatenate $\hat{F}^A_{1/8}$ and $F^A_{1/8}$ along the channel dimension. Then, we apply a point-wise convolution, decreasing the channel size to be equal to the channel size of $F^A_{1/4}$, followed by a depth-wise convolution operation with a kernel size of 3×3. The same procedure is applied to $\hat{F}^B_{1/8}$ and $F^B_{1/8}$. More formally:

$$\bar{F}^*_{1/8} = \mathrm{Conv}_{3\times3}\left(\mathrm{Conv}_{1\times1}(\hat{F}^*_{1/8} \parallel F^*_{1/8})\right), \tag{4}$$

where $*$ is either A or B. For each coarse match $(\tilde{i}, \tilde{j})$, we crop pairs of local windows at corresponding locations from $\{\bar{F}^A_{1/8}, \bar{F}^B_{1/8}\}$, $\{F^A_{1/4}, F^B_{1/4}\}$ and $\{F^A_{1/2}, F^B_{1/2}\}$ with sizes (1×1), (3×3) and (5×5) respectively. For one match pair $(\tilde{i}, \tilde{j})$, these windows are denoted as $\{f^A_{1\times1}, f^B_{1\times1}\}$, $\{f^A_{3\times3}, f^B_{3\times3}\}$ and $\{f^A_{5\times5}, f^B_{5\times5}\}$. If an index i or j is observed more than once in matches, the corresponding window is copied more than once for each pair $(\tilde{i}, \tilde{j})$. Next, to pass information between local window layers, we downsample the channel dimension of $f^A_{1\times1}$ and concatenate with $f^A_{3\times3}$, and then pass it through a transformer layer with a self-attention. From the output of the transformer, the windows are splinted back and denoted as $\bar{f}^A_{1\times1}$ and $\bar{f}^A_{3\times3}$. These steps can be formulated as:

$$\{\bar{f}^A_{1\times1}, \bar{f}^A_{3\times3}\} = \mathrm{Split}\left(\mathrm{Tr}_{\mathrm{self}}\left(\mathrm{Cat}(\mathrm{Down}(f^A_{1\times1}), f^A_{3\times3})\right)\right), \tag{5}$$

where Cat and Split are token-wise concatenation and splitting operation. $\mathrm{Tr}_{\mathrm{self}}$ is a transformer layer with self-attention, and Down denotes downsampling along the channel dimension. The same procedure is applied to $f^B_{1\times1}$ and $f^B_{3\times3}$ as well with outputs $\bar{f}^B_{1\times1}$ and $\bar{f}^B_{3\times3}$. After this step, to pass information across images, we use another transformer layer with cross-attention between $\bar{f}^A_{3\times3}$ and $\bar{f}^B_{3\times3}$, where the outputs are denoted as $\tilde{f}^A_{3\times3}$ and $\tilde{f}^B_{3\times3}$ as expressed by $\{\tilde{f}^A_{3\times3}, \tilde{f}^B_{3\times3}\} = \mathrm{Tr}_{\mathrm{cross}}(\bar{f}^A_{3\times3}, \bar{f}^B_{3\times3})$ where $\mathrm{Tr}_{\mathrm{cross}}$ is transformer layer with cross-attention.

Next, we apply the same steps from the start for window pairs $\{\tilde{f}^A_{3\times3}, f^A_{5\times5}\}$ and $\{\tilde{f}^B_{3\times3}, f^B_{5\times5}\}$, and obtain the outputs $\tilde{f}^A_{5\times5}$ and $\tilde{f}^B_{5\times5}$. For every coarse match prediction $(\tilde{i}, \tilde{j})$, the similarity matrix $\mathcal{S}_f^{\tilde{i},\tilde{j}}$ between fine-level windows $\tilde{f}^A_{5\times5}$ and $\tilde{f}^B_{5\times5}$ is calculated by $\mathcal{S}_f^{\tilde{i},\tilde{j}}(i,j) = \frac{1}{\tau} \cdot \langle \tilde{f}^A_{5\times5}(i), \tilde{f}^B_{5\times5}(j) \rangle$. Then, we employ dual-softmax operation to obtain fine-level matching probability matrix $\mathcal{P}_f^{\tilde{i},\tilde{j}}$:

$$\mathcal{P}_f^{\tilde{i},\tilde{j}}(i,j) = \mathrm{Softmax}\left(\mathcal{S}_f^{\tilde{i},\tilde{j}}(i,\cdot)\right)_j \cdot \mathrm{Softmax}\left(\mathcal{S}_f^{\tilde{i},\tilde{j}}(\cdot,j)\right)_i. \tag{6}$$

Finally, for each coarse match prediction $(\tilde{i}, \tilde{j})$, we select the pairs $(\hat{i}, \hat{j})$ for which $\mathcal{P}_f^{\tilde{i},\tilde{j}}(i,j)$ is higher than a threshold of θ_f and all other elements to obtain fine-level match predictions $\mathcal{M}_f$. As a result for each coarse-level match prediction $(\tilde{i}, \tilde{j})$, we constructed a fine-level match prediction $(\hat{i}, \hat{j})$.

In the employed transformer architectures, we use vanilla attention [74] and bidirectional attention [49, 77] for self and cross-attention layers respectively making it more robust to input variations without increasing computational complexity due to small window size and shared query and key projections. Furthermore, we add absolute positional bias to each window feature before sending it to transformer layers to leverage position information effectively. Inspired from [32, 53], we utilize a 2-layer MLP to embed the absolute 2D token location into the feature dimension.

3.3. Sub-pixel Refinement Module

In this module, we convert fine-level match predictions to sub-pixel matches by defining a simple regression mechanism on matches. In contrast to [79], we regress pixel locations for both images. For this purpose, we concatenate the feature vectors of $\tilde{f}^A_{5\times5}$ and $\tilde{f}^B_{5\times5}$ at $(\hat{i}, \hat{j})$ and apply MLP layer and Tanh function to jointly regress local sub-pixel coordinates $\delta_{x_A}, \delta_{y_A}, \delta_{x_B}, \delta_{y_B}$ as follows:

$$\{\delta_{x_A}, \delta_{y_A}, \delta_{x_B}, \delta_{y_B}\} = \mathrm{Tanh}\left(\mathrm{MLP}(\tilde{f}^A_{5\times5}(\hat{i}) \parallel \tilde{f}^B_{5\times5}(\hat{j}))\right). \quad (7)$$

Then, sub-pixel matches $(\hat{x}_A, \hat{x}_B) \in M_{sub}$ are obtained by summing local sub-pixel coordinates and coordinates of fine-level matches on the images. The sequence of fine-level matching followed by sub-pixel refinement for each coarse match allows us to prevent one point in image I^A from being matched to more than one point in the other image I^B and vice versa.

3.4. Masked Image Modeling

Before learning to match RGB-IR images, we introduce our model to real multi-modal image pairs with non-linear intensity differences belonging to visible and thermal spectra by utilizing MIM pre-training. Inspired by Pmatch [89], we pre-train our network to reconstruct randomly masked visible-thermal image pairs while conveying pre-trained encoder and decoder layers together to the fine-tuning task.

Masking Strategy: To use different scale feature maps in the encoder layer used in FLMM, we create the mask in the fine-scale and upscale it up to the original image resolution as in ConvNextv2 [78]. For both images, the masks are generated randomly to cover 50% of the image with 64×64 size patches. Instead of 32×32 as in [78, 81, 89], we use 64×64 patches with a larger input image size of 640×410 to enable the network to learn intensity differences in thermal and visible spectra in more detail. We start the masking procedure by applying the binary masks directly on the input images to avoid leakage of masked patches. After passing through the CNN backbone, similar to [6, 81, 89], we use learnable masked token vectors to replace the masked patches on feature maps $F^A_{1/8}$, $F^A_{1/4}$, $F^A_{1/2}$.

Decoder: After the LoFTR module, we directly employ the decoder architecture in FLMM to reconstruct the images. To implement this in practice, for each masked token in coarse scale, we create the local window $\tilde{f}_{5\times5}$ as described in FLMM section and then project it to original image resolution by $\tilde{I}_{10\times10} = \mathrm{Linear}(\tilde{\mathrm{f}}_{5\times5})$ to reconstruct the image. In other words, we reconstruct the image using 10×10 image windows for each coarse-level masked token. Thanks to the low disparity in the selected dataset [39], the same location of the feature maps for both images are used with the FLMM layer to benefit from the cross-attention layer. To supervise MIM, we use mean square error (MSE) between the target image and reconstructed image similar to [34, 78]. A sample masked sample images and their reconstruction results are shown in Fig. 3.

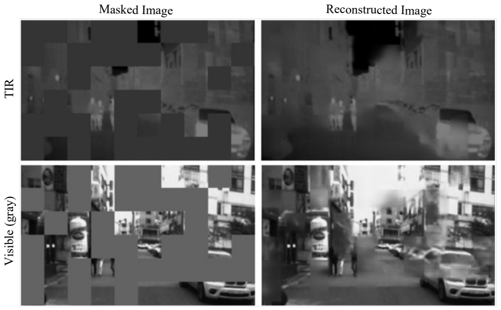

Figure 3. Visualization of reconstructed images using MIM pretext task. Input images are from [39].

3.5. Data Augmentation

Due to the lack of urban visible-thermal datasets to be used in image matching, we propose a simple but highly effective image augmentation method to generate pseudo-thermal images from visible images during the fine-tuning stage. To create a pseudo-thermal image, first, we randomly change the hue, saturation, and value intensities of the visible band RGB image. After converting the image to the grayscale, we apply a modified version of the cosine transformation [27, 83] to generate a variety of randomly generated images to represent thermal images with different intensity differences. For the grayscale image $I_{ij} \in [0, 1]$, the randomized cosine transformation is calculated by:

$$\begin{aligned} I_{pseudo} &= \mathrm{Norm}\left(\cos\left(\bar{w} \times (I - 0.5) + \bar{\theta}\right)\right), \text{ for} \\ \bar{w} &= w_0 + |\alpha_0| \times w_r, \\ \bar{\theta} &= \pi/2 + \alpha_1 \times \theta_r, \end{aligned} \quad (8)$$

where Norm stands for the min-max normalization of the image between 0 and 1. α_0 and α_1 are random variables with Normal distribution. w_0, w_r and θ_r are hyperparameters chosen as $2\pi/3$, $\pi/2$ and $\pi/2$ intuitively. We additionally apply random Gaussian blur operation with kernel size 5×5. Some generated pseudo-thermal image samples are shown in Fig. 4 together with real counterparts. By practicing the proposed augmentation method to one of the image pairs during fine-tuning, our network gains endurance against nonlinear intensity variations which is the crucial part for visible-thermal image matching.

3.6. Supervision

Our loss function consists of three components which are coarse-level matching loss, fine-level matching loss and sub-pixel refinement loss.

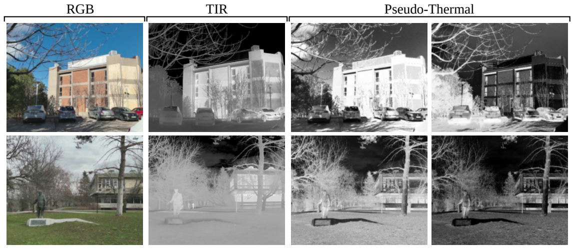

Figure 4. Pseudo-thermal image samples generated with the proposed augmentation method together with real counterparts.

Coarse-Level Matching Loss: To supervise the matching probability matrices P^0 and P^1, we apply the Focal Loss (FL) [48] following LoFTR [69] and AdaMatcher [38]:

$$\mathcal{L}_c = \left(FL(P^0, \hat{P}) + FL(P^1, \hat{P}) \right), \tag{9}$$

where $\hat{P}$ is coarse-level ground-truth matching matrix. We obtain ground-truth coarse matches similar to LoFTR [69] but without a mutual nearest neighbor constraint. For this purpose, we create 2D position grids for each image at the $1/8$ scale, and we project these grids to each other using depth maps and camera poses. Then, we assign projected grid points as positive matches using re-projection distances allowing one-to-many and many-to-one assignments.

Fine-Level Matching Loss: Although we select only one point as a match from fine-level windows $\tilde{f}^A_{5\times5}$ and $\tilde{f}^B_{5\times5}$, we supervise all fine-level features correspondences in $\mathcal{P}^{\tilde{i},\tilde{j}}_f$. We define fine-level matching loss as follows:

$$\mathcal{L}_f = \frac{1}{|\mathcal{M}_c|} \sum_{(\tilde{i},\tilde{j})\in\mathcal{M}_c} FL\left(\mathcal{P}^{\tilde{i},\tilde{j}}_f, \hat{\mathcal{P}}^{\tilde{i},\tilde{j}}_f\right), \tag{10}$$

where $\hat{\mathcal{P}}^{\tilde{i},\tilde{j}}$ is fine-level ground-truth matching matrix for a coarse-level match $(\tilde{i},\tilde{j})$. $\hat{\mathcal{P}}^{\tilde{i},\tilde{j}}$ is calculated at $1/2$ scale similar to coarse-level ground-truth matching matrix with an addition of mutual nearest neighbor constraint allowing only one-to-one matches.

Sub-pixel Refinement Loss: Inspired by TopicFM+ [28], we implemented the symmetric epipolar distance function [31] to calculate the sub-pixel refinement loss. This approach eliminates the need for explicit ground-truth matching pairs and enables us to supervise both matching coordinates jointly. Given an estimated matching coordinate pair $(\hat{x}_A, \hat{x}_B)$ in normalized image coordinates (in homogeneous form), the sub-pixel refinement loss is defined as:

$$\mathcal{L}_{sub} = \frac{1}{|\mathcal{M}_c|} \sum_{(\hat{x}_A,\hat{x}_B)} \|\hat{x}_A^T E \hat{x}_B\|^2 \left(\frac{1}{\|E^T\hat{x}_A\|^2_{0:2}} + \frac{1}{\|E\hat{x}_B\|^2_{0:2}} \right), \tag{11}$$

where E is the ground-truth essential matrix obtained using camera poses.

Overall Loss: Our total loss is calculated by:

$$\mathcal{L}_{total} = \lambda_c \mathcal{L}_c + \lambda_f \mathcal{L}_f + \lambda_{sub} \mathcal{L}_{sub}, \tag{12}$$

where λ_c, λ_f and λ_{sub} are hyperparameters chosen as 0.5, 0.3 and 10^4 respectively.

4. Proposed Dataset

To showcase the effectiveness of our method XoFTR, we introduce METU-VisTIR, a novel dataset featuring thermal and visible images captured across six diverse scenes with ground-truth camera poses. Four of the scenes encompass images captured under both cloudy and sunny conditions, while the remaining two scenes exclusively feature cloudy conditions. This diverse dataset facilitates the evaluation of matching algorithms across various challenges, including extreme viewpoint variations and weather-induced changes in lighting and temperature.

We captured sequential images of the scenes using cameras of DJI Mavic 3 Thermal drone whose thermal and visible band cameras are positioned closely. The thermal camera boasts a resolution of 640×512 pixels, a FOV of $61°$, and operates within the wavelength range of 8-14 μm. Meanwhile, the visible band RGB camera features a resolution of 3840×2160 pixels and a FOV of $84°$. Ground-truth poses were recovered using offline systems such as COLMAP [65, 67] and HLOC [62] methods with RGB images. Since the cameras are auto-focus, we obtained GT camera parameters for both cameras using COLMAP in the same modality, supplemented by distortion parameter estimation using a calibration pattern. Although the auto-focus nature of the cameras may result in slight imperfections in the ground truth camera parameters, they are adequate for the purpose of method evaluation. Some images from our dataset are shown in Fig. 5.

We created two benchmark sets from the captured images, totaling 1382 and 1208 image pairs, labeled as cloudy-cloudy and cloudy-sunny. The cloudy-cloudy set consists of thermal and RGB image pairs with corresponding GT camera poses, all captured under cloudy conditions. Conversely, the cloudy-sunny set contains thermal and RGB image pairs with GT poses, capturing one image in sunny and the other in cloudy conditions.

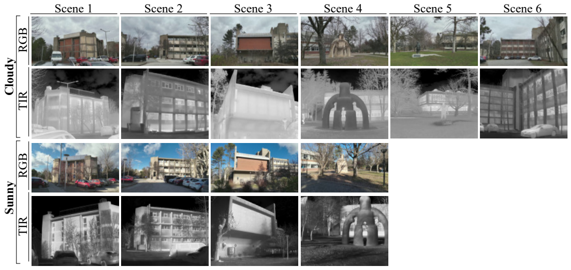

Figure 5. Visualization of some images from our dataset.

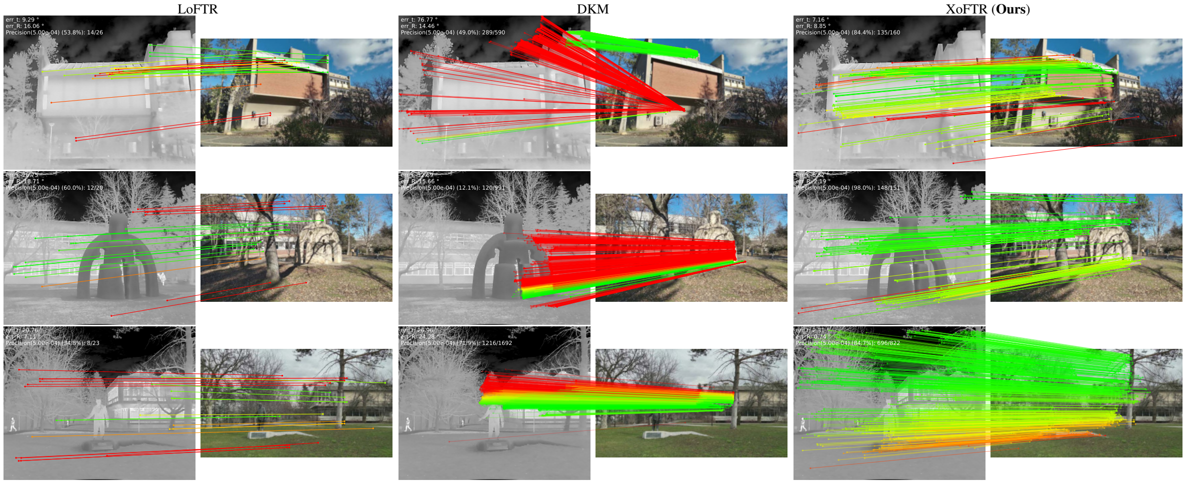
LoFTR
DKM
XoFTR (Ours)
err_t: 9.29 °
err_R: 16.06 °
Precision(5.00e-04) (53.8%): 14/26
err_t: 76.77 °
err_R: 14.46 °
Precision(5.00e-04) (49.0%): 289/590
err_t: 7.16 °
err_R: 8.85 °
Precision(5.00e-04) (84.4%): 135/160

Figure 6. **Qualitative results for pose estimation.** XoFTR (right column) is compared to DKM and LoFTR in METU-VisTIR dataset. Only the inlier matches after RANSAC are shown, and matches with epipolar error below 5×10^{-4} are shown in green lines.

Category	Method	Pose estimation AUC		
		@5°	@10°	@20°
Detector-based	D2-Net [25]+NN	2.16	6.01	12.80
	SP [20]+SuperGlue [63]	3.90	8.75	16.35
	SP [20]+LightGlue [49]	1.17	3.97	9.60
	ReDFeat [18]	2.36	5.45	11.26
Detector-free	LoFTR [69]	2.63	6.55	14.11
	LoFTR-MTV [52]	1.54	3.89	8.80
	ASpanFormer [14]	1.82	4.73	10.60
	DKM [26]	<u>5.79</u>	<u>11.47</u>	<u>19.17</u>
	XoFTR (**Ours**)	**22.03**	**39.03**	**55.06**

Table 2. **Evaluation on METU-VisTIR cloudy-cloudy dataset.** Relative pose estimation results for visible-thermal image pairs taken under cloudy weather conditions.

Category	Method	Pose estimation AUC		
		@5°	@10°	@20°
Detector-based	D2-Net [25]+NN	1.13	3.66	8.96
	SP [20]+SuperGlue [63]	4.06	10.70	19.91
	SP [20]+LightGlue [49]	3.12	8.39	15.29
	ReDFeat [18]	1.16	3.24	7.22
Detector-free	LoFTR [69]	2.77	7.71	16.36
	LoFTR-MTV [52]	0.92	2.79	6.82
	ASpanFormer [14]	3.18	7.13	14.01
	DKM [26]	<u>7.26</u>	<u>14.63</u>	<u>23.60</u>
	XoFTR (**Ours**)	**12.59**	**27.90**	**45.03**

Table 3. **Evaluation on METU-VisTIR cloudy-sunny dataset.** Relative pose estimation results for visible-thermal images taken under the cloudy and the sunny weather conditions.

5. Experiments

5.1. Implementation Details

We pre-train our model on the KAIST Multispectral Pedestrian Detection [39] dataset, containing 95,000 visible-thermal pairs from a moving vehicle, with a focus on the top 640×480 region to avoid road-dominant lower parts. We use Adam for pre-training with an initial learning rate of 2.5×10^{-4} and a batch size of 2 for 9 epochs, taking 24 hours on 2 A5000 GPUs. For fine-tuning, we use the MegaDepth [46] dataset with a 16 batch size at 640 × 640 resolution for padded images, employing Adam with a 2×10^{-3} learning rate, converging after 24 hours on 8 A100 GPUs. Augmentation is applied randomly to one of the images I^A or I^B. θ_c, θ_f thresholds: 0.3, 0.1 respectively.

5.2. Experiment 1: Relative Pose Estimation

Evaluation protocol: To evaluate our method with the METU-VisTIR, following [69], we assess pose error using area under curve (AUC) at 5°, 10°, and 20° thresholds, defined as the maximum angular deviation from GT in rotation and translation. We employ RANSAC and a 1.5 threshold to solve for the essential matrix with predicted matches, setting the longer image side to 640 pixels during testing. Evaluations were independently conducted for cloudy-cloudy and cloudy-sunny sets.

Compared methods: We compared our XoFTR with the following publicly available methods: (1) Detector-based methods including D2-Net [25], SuperGlue [63], LightGlue [49] and ReDFeat [18] , and (2) detector-free matchers including LoFTR [69], LoFTR-MTV [52], ASpanFormer [14] and DKM [26]. LoFTR-MTV [52] is LoFTR trained with aerial visible-TIR image pairs. Prior work in multi-modal image matching often lack public code or aren't readily benchmarked as multi-modal baselines.

Results: As shown in Tab. 2 and 3, XoFTR outperforms the other methods by a large margin in terms of relative pose estimation, demonstrating XoFTR's effectiveness. When examining the LoFTR-MTV model trained on real **aerial** thermal and visible band images, it becomes evident that the proposed training approach is crucial for achieving accurate matching in urban settings. When we compare Tab.

Category	Method	Homography est. AUC		
		@5°	@10°	@20°
Detector-based	D2-Net [25]+NN	2.17	6.10	16.85
	SP [20]+SuperGlue [63]	4.76	15.99	37.95
	SP [20]+LightGlue [49]	5.57	15.83	35.42
	ReDFeat [18]	3.72	12.13	29.21
Detector-free	LoFTR [69]	6.34	14.22	30.23
	LoFTR-MTV [52]	4.56	8.57	16.23
	ASpanFormer [14]	**9.50**	<u>18.87</u>	<u>36.42</u>
	DKM [26]	2.79	9.55	25.87
	XoFTR **(Ours)**	<u>8.19</u>	**23.37**	**48.15**

Table 4. **Homography estimation on LGHD LWIR/RGB [2] and FusionDN [82] datasets.** The AUC of the corner error is reported in percentage.

2 and 3, we observe a decrease in the performance of our method due to the varying weather conditions. The qualitative results in Fig. 6 support the quantitative results.

5.3. Experiment 2: Homography Estimation

Dataset: We utilized the LGHD LWIR/RGB [2] dataset along with the RoadScene dataset from FusionDN [82], which contains 221 pairs of aligned visible-infrared images of road scenes with vehicles and pedestrians. The LGHD LWIR/RGB dataset comprises 44 pairs of aligned visible-thermal images of buildings. By merging these datasets, we obtained a new dataset. For each image pair, we generated 5 different homographies to serve as ground truth (GT) and applied them to the images, yielding a total of 1325 image pairs. Generated GT homographies include random scaling of $[0.8, 1.2]$, random perspective distortion $[-0.15, +0.15]$, and random rotation $[-15, +15]$ degrees.

Evaluation protocol: We used the same evaluation protocol that LoFTR uses for the HPatches [5] dataset, presenting results in terms of the area under the cumulative curve (AUC) for corner error distances of 3, 5, and 10 pixels.

Results: Tab. 4 demonstrates that XoFTR surpasses other baseline methods across for 10 and 20 pixel error thresholds by a notable margin while Aspanformer [14] gets the best result for the 5-pixel error threshold. Notably, the performance disparity between LoFTR and alternative approaches widens as the correctness threshold increases. This experiment confirms that our model performs well beyond our dataset, demonstrating its success in various situations. For qualitative comparison, see Fig. 7.

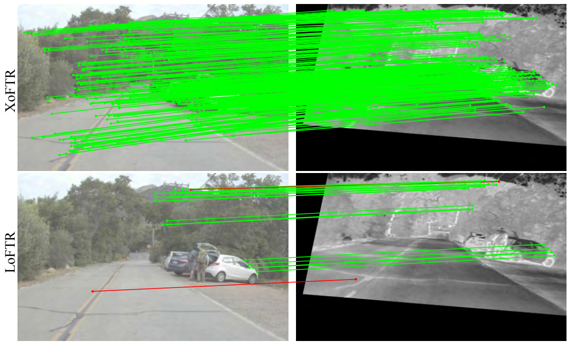

Figure 7. The qualitative homography estimation results for XoFTR and LoFTR [69].

Method	Pose estimation AUC		
	@5°	@10°	@20°
(1) without pretraining	11.81	26.51	42.93
(2) without augmentation	2.92	7.43	14.94
(3) with only one-to-one assignment	9.95	22.60	38.73
(4) with coarse-to-fine module of LoFTR	6.31	14.46	26.77
(5) without SPRM	12.31	27.39	44.68
(6) without the second thresholding θ_f	11.58	26.08	43.56
(7) without the positional bias in FLMM	12.23	26.99	43.36
Full (XoFTR)	**12.59**	**27.90**	**45.03**

Table 5. **Ablation study of XoFTR.** All variants of XoFTR are evaluated on METU-VisTIR cloud-sunny for pose estimation.

5.4. Experiment 3: Ablation Study

We assess five different variants of XoFTR evaluated on our METU-VisTIR cloudy-sunny dataset (Tab. 5). The results suggest that: (1) Training from scratch for only image matching w/o pretext task yields an AUC drop as expected. (2) Fine-tuning w/o augmentation leads to a significant drop in AUC, showing the effectiveness of our proposed augmentation method. (3) Allowing only one-to-one assignment in coarse matching as in LoFTR results in a considerable drop in AUC, demonstrating the importance of one-to-many assignment. (4) Replacing our FLLM and CLMM with LoFTR's coarse-to-fine module (with one-to-one assignment) leads to a serious drop in AUC, showing the effects of the methods we use when bringing coarse matches to sub-pixel resolution. (5) Matching at the $1/2$ scale w/o SPRM yields an AUC drop due to imprecise matches. (6) Removing the second thresholding (θ_f), which filters low confidence matches at $1/2$ scale, lowers AUC. (7) Removing the absolute positional bias from FLMM results in a drop in AUC.

Running Time On an A5000 GPU, XoFTR runs 116 ms at 640 × 512 resolution while LoFTR [69] runs 102 ms: One-to-many assignment increases the number of the coarse matches leading to a small amount of process time increase.

6. Conclusion

We have introduced XoFTR as a novel pipeline for cross-view visible-thermal image matching. Our two-stage approach significantly outperforms the compared methods on several benchmarks. To better evaluate methods, we have also introduced a novel challenging dataset.

Acknowledgements: This work is funded by ROKETSAN Inc. The numerical calculations were partially performed at TÜBITAK ULAKBIM, TRUBA.

References

[1] Florian Achermann, Andrey Kolobov, Debadeepta Dey, Timo Hinzmann, Jen Jen Chung, Roland Siegwart, and Nicholas Lawrance. Multipoint: Cross-spectral registration of thermal and optical aerial imagery. In *Conference on Robot Learning*, pages 1746–1760. PMLR, 2021. 1, 2

[2] Cristhian Aguilera, Angel D. Sappa, and Ricardo Toledo. Lghd: A feature descriptor for matching across non-linear intensity variations. In *Image Processing (ICIP), 2015 IEEE International Conference on*, page 5. IEEE, 2015. 8

[3] Moab Arar, Yiftach Ginger, Dov Danon, Amit H Bermano, and Daniel Cohen-Or. Unsupervised multi-modal image registration via geometry preserving image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13410–13419, 2020. 2

[4] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders masked inputs multimae predictions target masked inputs multimae predictions target masked inputs multimae predictions target, 2023. 3

[5] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *CVPR*, 2017. 8

[6] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *ICLR 2022 - 10th International Conference on Learning Representations*, 2022. 2, 5

[7] Elad Ben Baruch and Yosi Keller. *Multimodal matching using a hybrid convolutional neural network*. PhD thesis, Ben-Gurion University of the Negev, 2018. 2

[8] Elad Ben Baruch and Yosi Keller. Joint detection and matching of feature points in multimodal images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 6585–6593, 2021. 1

[9] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*, pages 404–417. Springer, 2006. 2

[10] Guillaume-Alexandre Bilodeau, Atousa Torabi, and François Morin. Visible and infrared image registration using trajectories and composite foreground images. *Image and Vision Computing*, 29(1):41–50, 2011. 1, 2

[11] Samarth Brahmbhatt, Jinwei Gu, Kihwan Kim, James Hays, and Jan Kautz. Geometry-aware learning of maps for camera localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2616–2625, 2018. 1

[12] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. Brief: Binary robust independent elementary features. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, pages 778–792. Springer, 2010. 2

[13] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2

[14] Hongkai Chen, Zixin Luo, Lei Zhou, Yurun Tian, Mingmin Zhen, Tian Fang, David Mckinnon, Yanghai Tsin, and Long Quan. Aspanformer: Detector-free image matching with adaptive span transformer. In *European Conference on Computer Vision*, pages 20–36. Springer, 2022. 1, 2, 7, 8

[15] Jiaxuan Chen, Xiaoxian Chen, Shuang Chen, Yuyan Liu, Yujing Rao, Yang Yang, Haifeng Wang, and Dan Wu. Shapeformer: Bridging cnn and transformer via shapeconv for multimodal image matching. *Information Fusion*, 91:445–457, 2023. 1, 2

[16] Yuxing Chen, Maofan Zhao, and Lorenzo Bruzzone. Incomplete multimodal learning for remote sensing data fusion. *arXiv preprint arXiv:2304.11381*, 2023. 3

[17] Song Cui, Ailong Ma, Yuting Wan, Yanfei Zhong, Bin Luo, and Miaozhong Xu. Cross-modality image matching network with modality-invariant feature representation for airborne-ground thermal infrared and visible datasets. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021. 1, 2

[18] Yuxin Deng and Jiayi Ma. Redfeat: Recoupling detection and description for multimodal feature learning. *IEEE Transactions on Image Processing*, 32:591–602, 2022. 2, 7, 8

[19] Bhavesh Deshpande, Sourabh Hanamsheth, Yawen Lu, and Guoyu Lu. Matching as color images: Thermal image local feature detection and description. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1905–1909. IEEE, 2021. 1, 3

[20] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 2, 7, 8

[21] Chaitanya Devaguptapu, Ninad Akolekar, Manuj M. Sharma, and Vineeth N. Balasubramanian. Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2019-June, 2019. 3

[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[23] Yide Di, Yun Liao, Kaijun Zhu, Hao Zhou, Yijia Zhang, Qing Duan, Junhui Liu, and Mingyu Lu. Mivi: Multi-stage feature matching for infrared and visible image. *The Visual Computer*, pages 1–13, 2023. 2

[24] Elan Dubrofsky. Homography estimation. *Diplomová práce. Vancouver: Univerzita Britské Kolumbie*, 5, 2009. 1

[25] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint description and detection of local features. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 8092–8101, 2019. 2, 7, 8

[26] Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. Dkm: Dense kernelized feature matching for geometry estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17765–17775, 2023. 2, 7, 8

[27] C. Fookes, A. Maeder, S. Sridharan, and J. Cook. Multi-spectral stereo image matching using mutual information. In *Proceedings. 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004.*, pages 961–968, 2004. 2, 3, 5

[28] Khang Truong Giang, Soohwan Song, and Sungho Jo. Topicfm+: Boosting accuracy and efficiency of topic-assisted feature matching, 2023. 6

[29] Jungong Han, Eric J Pauwels, and Paul De Zeeuw. Visible and infrared image registration in man-made environments employing hybrid visual features. *Pattern Recognition Letters*, 34(1):42–51, 2013. 1, 2

[30] Junlin Han, Mehrdad Shoeiby, Lars Petersson, and Mohammad Ali Armin. Dual contrastive learning for unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 3

[31] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 6

[32] Ali Hatamizadeh, Greg Heinrich, Hongxu Yin, Andrew Tao, Jose M Alvarez, Jan Kautz, and Pavlo Molchanov. Fastervit: Fast vision transformers with hierarchical attention. *arXiv preprint arXiv:2306.06189*, 2023. 5

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[34] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2022. 2, 5

[35] Zhuolu Hou, Yuxuan Liu, and Li Zhang. Pos-gift: A geometric and intensity-invariant feature transformation for multimodal images. *Information Fusion*, 102:102027, 2024. 1

[36] Tomislav Hrkać, Zoran Kalafatić, and Josip Krapac. Infrared-visual image registration based on corners and hausdorff distance. In *Image Analysis: 15th Scandinavian Conference, SCIA 2007, Aalborg, Denmark, June 10-14, 2007 15*, pages 383–392. Springer, 2007. 2

[37] Maoqing Hu, Bin Sun, Xudong Kang, and Shutao Li. Multiscale structural feature transform for multi-modal image matching. *Information Fusion*, 95:341–354, 2023. 1

[38] Dihe Huang, Ying Chen, Yong Liu, Jianlin Liu, Shang Xu, Wenlong Wu, Yikang Ding, Fan Tang, and Chengjie Wang. Adaptive assignment for geometry aware local feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5425–5434, 2023. 1, 2, 3, 6

[39] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1037–1045, 2015. 5, 7

[40] Xingyu Jiang, Jiayi Ma, Guobao Xiao, Zhenfeng Shao, and Xiaojie Guo. A review of multimodal image matching: Methods and applications. *Information Fusion*, 73:22–71, 2021. 1

[41] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pages 5156–5165. PMLR, 2020. 3

[42] Vladimir V. Kniaz, Vladimir A. Knyaz, Jiří Hladůvka, Walter G. Kropatsch, and Vladimir A. Mizginov. ThermalGAN: Multimodal Color-to-Thermal Image Translation for Person Re-Identification in Multispectral Dataset. In *Computer Vision – ECCV 2018 Workshops*. Springer International Publishing, 2018. 3

[43] Dong-Guw Lee, Myung-Hwan Jeon, Younggun Cho, and Ayoung Kim. Edge-guided multi-domain rgb-to-tir image translation for training vision tasks with challenging labels. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8291–8298. IEEE, 2023. 3

[44] Hsin Ying Lee, Hung Yu Tseng, Qi Mao, Jia Bin Huang, Yu Ding Lu, Maneesh Singh, and Ming Hsuan Yang. Drit++: Diverse image-to-image translation via disentangled representations. *International Journal of Computer Vision*, 128, 2020. 3

[45] Jiayuan Li, Qingwu Hu, and Mingyao Ai. Rift: Multi-modal image matching based on radiation-variation insensitive feature transform. *IEEE Transactions on Image Processing*, 29: 3296–3310, 2019. 1

[46] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 1, 7

[47] Zhao-Liang Li, Hua Wu, Ning Wang, Shi Qiu, José A Sobrino, Zhengming Wan, Bo-Hui Tang, and Guangjian Yan. Land surface emissivity retrieval from satellite data. *International Journal of Remote Sensing*, 34(9-10):3084–3127, 2013. 1

[48] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 6

[49] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17627–17638, 2023. 1, 2, 4, 7, 8

[50] Xiangzeng Liu, Yunfeng Ai, Bin Tian, and Dongpu Cao. Robust and fast registration of infrared and visible images for electro-optical pod. *IEEE Transactions on Industrial Electronics*, 66(2):1335–1344, 2018. 2

[51] Xiangzeng Liu, Yunfeng Ai, Juli Zhang, and Zhuping Wang. A novel affine and contrast invariant descriptor for infrared and visible image registration. *Remote Sensing*, 10(4):658, 2018. 2

[52] Yuxiang Liu, Yu Liu, Shen Yan, Chen Chen, Jikun Zhong, Yang Peng, and Maojun Zhang. A multi-view thermal–visible image dataset for cross-spectral matching. *Remote Sensing*, 15(1):174, 2022. 7, 8

[53] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 5

[54] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 2

[55] Aviad Moreshet and Yosi Keller. Attention-based multimodal image matching. *Computer Vision and Image Understanding*, page 103949, 2024. 1, 2

[56] Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, 2020. 3

[57] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 2

[58] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *OpenAI*, 2018. 2

[59] Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. R2d2: Reliable and repeatable detector and descriptor. *Advances in neural information processing systems*, 32, 2019. 2

[60] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. *Advances in neural information processing systems*, 31, 2018. 2

[61] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011. 2

[62] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *CVPR*, 2019. 6

[63] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 2, 7, 8

[64] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, et al. Benchmarking 6dof outdoor visual localization in changing conditions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8601–8610, 2018. 1

[65] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6

[66] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 1

[67] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 6

[68] Ukcheol Shin, Kyunghyun Lee, and In So Kweon. Complementary random masking for rgb-thermal semantic segmentation, 2023. 3

[69] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 1, 2, 3, 6, 7, 8

[70] Yurun Tian, Bin Fan, and Fuchao Wu. L2-net: Deep learning of discriminative patch descriptor in euclidean space. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 661–669, 2017. 2

[71] Fabio Tosi, Pierluigi Zama Ramirez, Matteo Poggi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. Rgb-multispectral matching: Dataset, learning methodology, evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15958–15968, 2022. 1

[72] Prune Truong, Martin Danelljan, Luc V Gool, and Radu Timofte. Gocor: Bringing globally optimized correspondence volumes into your neural network. *Advances in Neural Information Processing Systems*, 33:14278–14290, 2020. 2

[73] Prune Truong, Martin Danelljan, and Radu Timofte. Glu-net: Global-local universal network for dense flow and correspondences. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6258–6268, 2020. 2

[74] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[75] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11, 2010. 2

[76] Lan Wang, Chenqiang Gao, Yue Zhao, Tiecheng Song, and Qi Feng. Infrared and visible image registration using transformer adversarial network. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1248–1252. IEEE, 2018. 2

[77] Phil Wang. Bidirectional cross attention. https://github.com/lucidrains/bidirectional-cross-attention, 2022. 4

[78] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16133–16142, 2023. 5

[79] Tao Xie, Kun Dai, Ke Wang, Ruifeng Li, and Lijun Zhao. Deepmatcher: a deep transformer-based network for robust and accurate local feature matching. *Expert Systems with Applications*, 237:121361, 2024. 5

[80] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 171–180. Springer, 2021. 2

[81] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: a simple framework for masked image modeling. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2022. 2, 5

[82] Han Xu, Jiayi Ma, Zhuliang Le, Junjun Jiang, and Xiaojie Guo. Fusiondn: A unified densely connected network for image fusion. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12484–12491, 2020. 8

[83] Mustafa Yaman and Sinan Kalkan. An iterative adaptive multi-modal stereo-vision method using mutual information. *Journal of Visual Communication and Image Representation*, 26:115–131, 2015. 2, 3, 5

[84] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 467–483. Springer, 2016. 2

[85] Lichao Zhang, Abel Gonzalez-Garcia, Joost Van De Weijer, Martin Danelljan, and Fahad Shahbaz Khan. Synthetic data generation for end-to-end thermal infrared tracking. *IEEE Transactions on Image Processing*, 28, 2019. 3

[86] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Ibot: Image bert pre-training with online tokenizer. In *ICLR 2022 - 10th International Conference on Learning Representations*, 2022. 2

[87] Qunjie Zhou, Torsten Sattler, and Laura Leal-Taixe. Patch2pix: Epipolar-guided pixel-level correspondences. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4669–4678, 2021. 2

[88] Zexian Zhou and Xiaojing Liu. Masked autoencoders in computer vision: A comprehensive survey. *IEEE Access*, 11:113560–113579, 2023. 2

[89] Shengjie Zhu and Xiaoming Liu. Pmatch: Paired masked image modeling for dense geometric matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21909–21918, 2023. 1, 2, 3, 5

[90] Mehmet Akif Özkanoğlu and Sedat Ozer. Infragan: A gan architecture to transfer visible images to infrared domain. *Pattern Recognition Letters*, 155, 2022. 3